In an interesting paper in this issue of ONKOLGIE, Edler and Kopp-Schneider [1] discuss a number of ‘myths’ concerning the principles involved in the statistical design and interpretation of the results of randomized phase III clinical trials. I would like to briefly comment on two of these myths.

The first myth which is rejected by the authors is that two independent positive studies are necessary to claim treatment efficacy. The authors suggest there may be instances where one study might be sufficient or that one could combine the results of two different studies. The Committee for Proprietary Medicinal Products (CPMP) has provided guidelines on this subject: Points to Consider on Validity and Interpretation of Meta-Analyses, and One Pivotal Study [2]. Although in most drug applications the proof of efficacy relies on several studies, the CPMP recognizes that there are cases where applications may be based on the results from one study or on the pooled results from several studies (meta-analyses). Although it is usually wise to plan more than one study in the phase III program of a new drug, there is no formal requirement to carry out two or more pivotal studies. The minimum requirement set forth by the CPMP [2] is one controlled study with statistically compelling and clinically relevant results. The following factors should be considered: the trial’s internal and external validity, clinical relevance, degree of statistical relevance, data quality, internal consistency, center effects, and the plausibility of the hypothesis tested. Thus, a single, high-quality, large, multi-center trial might provide sufficient evidence of treatment efficacy if the treatment results are statistically compelling, clinically relevant, and reasonably homogeneous across both institutions and prognostic factors.

Although meta-analyses may be problematic for a number of reasons, the CPMP has also spelled out the conditions under which meta-analyses may provide pivotal evidence [2]. While meta-analyses should preferably be prospectively defined, prerequisites for a potentially acceptable retrospective meta-analysis include: some studies clearly positive, inconclusive studies showing positive trends, no statistical or major heterogeneity, pooled confidence interval well away from zero, a justification that a biased selection of studies and/or endpoints is unlikely, and a sensitivity analysis demonstrating the robustness of the findings. They specifically state that a retrospective meta-analysis of only two studies originally intended to stand on their own will not generally be useful and that a meta-analysis cannot be used to reconcile the results of two conflicting studies, one which is positive and one which is negative. The clinical relevance of the findings, and the heterogeneity and external validity of the results must also be taken into account.

Properly done meta-analyses based on individual patient data from all randomized clinical trials dealing with a specific treatment comparison provide, in fact, the highest level of evidence-based medicine. They yield the best overall estimate of the size of the treatment effect, a confidence interval for the size of the difference, along with an assessment of the heterogeneity of results between the studies and eventually across the levels of different prognostic factors. Two examples are the adjuvant tamoxifen meta-analysis in early breast cancer [3] and the more recent meta-analysis of the effect of bacillus Calmette-Guerin on progression in superficial bladder cancer [4]. Few of the individual trials in either meta-analysis showed a significant treatment benefit, but the overall results from each meta-analysis provided quite striking differences in treatment efficacy which have had an important impact on clinical practice. Thus it is important to properly assess all the available evidence at hand and not to concentrate on whether one or two trials may or may not show statistically significant differences.

The other myth that I would like to comment on is that of ‘one hypothesis, one trial design’. The example given is a 3-arm trial where arm A is a new treatment, arm B is the standard treatment in standard duration, and arm C is the standard treatment in shorter duration. This example raises the question of the appropriate significance levels to use when faced with multiple testing.
More generally, multiple testing may take the form of multiple endpoints (primary and secondary), multiple treatment comparisons (more than two treatment groups), multiple statistical testing on the same endpoint (using several different statistical techniques, comparing survival at one, two, and five years), subgroup analyses (prognostic factors, intent to treat/per protocol analyses), safety analyses (multiple parameters), repeated significance testing over time (interim analyses), and even multiple trials. In all these cases it is important to take into account the possible impact of multiple comparisons, that is, the more comparisons are made, the more likely it is to reject by chance at least one null hypothesis even when all the null hypotheses are true (an inflated rate of false positive conclusions). Thus, the size of the type I error to be used for each comparison may need to be adjusted in order to maintain an overall study-wise type I error rate of 5% or less [5].

There are various methods that take multiplicity into account, reduce it, or avoid it. Once again, the CPMP has provided guidelines on this topic: Points to Consider on Multiplicity Issues in Clinical Trials [6]. There are two main situations when no formal adjustment of the type I error rate is required (closed testing procedures):

1. statistical significance is required at the alpha (5%) significance level for all the comparisons related to the trial's objectives (increases the type II error),
2. two or more comparisons are a priori ordered according to clinical relevance.

However, in this last case, no claims of efficacy can be made based on comparisons that have a rank lower than the first comparison whose null hypothesis could not be rejected. The hierarchical a priori order for testing the null hypotheses must be pre-specified in the protocol.

Another solution, when more flexibility is desired concerning the hypotheses to be tested, is to carry out each individual comparison at a more extreme significance level than 5% in order to maintain the overall type I error at 5% – for example 5% divided by the number of comparisons. This is the Bonferroni adjustment, which is known to be conservative in nature and hence is associated with a loss in power [7, 8].

A number of improvements to the Bonferroni adjustment procedure have been published, including the modified Bonferroni step up and step down procedures where the null hypotheses have been a posteriori ordered based on the corresponding ordered observed p values [8, 9]. In the step down (up) procedure, significance testing continues until one accepts (rejects) a null hypothesis. Then one accepts (rejects) all remaining (ordered, untested) null hypotheses without further testing.

With respect to the example with three treatment groups, the comparison of A to B is one of superiority (difference) while that of B to C is one of equivalence, or rather non-inferiority since it is desired to show that C is not worse than B. Equivalence trials (two sided) are quite rare in cancer whereas non-inferiority trials (one sided) are not. In each case the null hypothesis is the opposite of what one is trying to prove. However, the difference Δ used in the two comparisons should not be the same. Since relatively small differences must be ruled out in non-inferiority trials, sample sizes are generally much larger in non-inferiority trials than in superiority studies. In order not to give back all of the benefit of the control group, the CPMP [10] recommends that the difference used in non-inferiority studies should be no more than one third to one half the size of the benefit found for the control treatment in a previous superiority study. It is also recommended to carry out non-inferiority tests at the one sided 2.5% level [11].

Rather than doing two separate two-arm studies with a common control arm, it is generally preferable to do a three-arm study as long as the overall significance level is maintained at the 5% level. However, one possible problem with three-arm studies is to get investigator agreement and patient consent to be randomized to any of the treatment arms. In a three-arm trial, any a priori ordering of the hypotheses that might be used must make sense. In the example used, the value of testing B and C for non-inferiority only if A has been shown to be superior to B is quite questionable. In fact, testing B and C for non-inferiority is probably only of interest if A has been not been shown to be superior to B. In many cases it may be preferable to employ a pure Bonferroni approach with no a priori or a posteriori ordering of the hypotheses or p values so as not to be restricted with respect to the hypotheses that can be tested.

Richard Sylvester, Brussels

References